applied sciences

MDPI

Article

Privacy Preservation and Analytical Utility of E-Learning Data Mashups in the Web of Data

Mercedes Rodriguez-Garcia [1], Antonio Balderas [2,*] and Juan Manuel Dodero [2]

[1] Departamento de Ingeniería en Automática, Electrónica, Arquitectura y Redes de Computadores, Universidad de Cádiz, 11519 Puerto Real, Spain; mercedes.rodriguez@uca.es
[2] Departamento de Ingeniería Informática, Universidad de Cádiz, 11519 Puerto Real, Spain; juanma.dodero@uca.es
* Correspondence: antonio.balderas@uca.es

Abstract: Virtual learning environments contain valuable data about students that can be correlated and analyzed to optimize learning. Modern learning environments based on data mashups that collect and integrate data from multiple sources are relevant for learning analytics systems because they provide insights into students' learning. However, data sets involved in mashups may contain personal information of sensitive nature that raises legitimate privacy concerns. Average privacy preservation methods are based on preemptive approaches that limit the published data in a mashup based on access control and authentication schemes. Such limitations may reduce the analytical utility of the data exposed to gain students' learning insights. In order to reconcile utility and privacy preservation of published data, this research proposes a new data mashup protocol capable of merging and *k*-anonymizing data sets in cloud-based learning environments without jeopardizing the analytical utility of the information. The implementation of the protocol is based on linked data so that data sets involved in the mashups are semantically described, thereby enabling their combination with relevant educational data sources. The *k*-anonymized data sets returned by the protocol still retain essential information for supporting general data exploration and statistical analysis tasks. The analytical and empirical evaluation shows that the proposed protocol prevents individuals' sensitive information from re-identifying.

Keywords: learning analytics; data mashup; data privacy; privacy-preserving data publishing; *k*-anonymity

Citation: Rodriguez-Garcia, M.; Balderas, A.; Dodero, J. M. Privacy Preservation and Analytical Utility of E-Learning Data Mashups in the Web of Data. *Appl. Sci.* **2021**, *11*, 8506. https://doi.org/10.3390/app11188506

Academic Editor: Gianluca Lax

Received: 19 August 2021
Accepted: 10 September 2021
Published: 13 September 2021

1. Introduction

Aware of data opportunities in an information-driven world, private, academic, and government organizations are including frameworks that enable the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) on Big Data Governance and Metadata Management in their roadmap [1]. In particular, the IEEE Standards Association states the need for devising interoperable data architectures that enable effective integration—i.e., *mashup*—of data from multiple sources to answer specific information requests [2]. Different data integration approaches from diverse application domains attempt to cover the requirement of interoperability by providing public data access infrastructures that enable dataset mashups [2], such as the Common Access Platform, proposed by the National Institute of Standards and Technology (NIST) [3–5]. By unifying information from diversified data repositories, companies and organizations can generate value-added information and, consequently, detect new business opportunities, identify risks, and discover new patterns and insights. A great variety of sectors can benefit from Big Data integration to empower their analytic systems, including social media and search engines; insurance, banking, and finances; marketing; retail and point-of-sale analytics; manufacturing optimization; transportation; utility and energy; healthcare; and research and development [6].

In educational institutions, large amounts of ubiquitous data about students, available from different cloud-based data sources [7], include students' demographics besides relevant data about students' learning. They can be merged with data from the institution's student records, digital libraries, and Learning Management Systems (LMS) to build customized Virtual Learning Environments (VLE) as personal learning mashups [8]. Thanks to the simplicity and low complexity of web standards, the first generation of personal learning mashups have been based on a composition of web-based educational services [9]. However, the automation of independent remote services orchestration is not straightforward, yielding an obstacle for implementing and using service-based learning mashups [10]. In contrast, new generation VLEs are based on mashing up data available in realistic cloud-based learning environments [7], which may even involve data from independent educational institutions [11]. Eventually, Learning Analytics (LA) systems built on new generation learning mashups can benefit from the fusion of large amounts of data gathered from cloud-based learning environments and institutional teaching systems to obtain new insights that can improve teachers' learning design practices [12] and students' learning performance [13].

Supporting data mashups is critical to assist in data-driven decision-making. However, sharing and mashing up information with personal content may compromise the privacy of individuals referenced in the data. Regulations on data privacy, such as the Family Educational Rights and Privacy Act (FERPA) [14], point out that shared data with unique identifiers removed can still lead to the re-identification of individuals through data linkage attacks [15] by correlating potentially identifying combinations of attributes—called Quasi-Identifiers (QI)—with publicly available external information. Consequently, and since the trust of the educational community is essential for the adoption of LA, there is a need to incorporate anonymization mechanisms in LA that guarantee information privacy [13]. More generally, the NIST identifies the balance between privacy and utility as one of the main issues to be addressed in interoperable data architectures [16].

The dual need to share information and protect privacy at the same time has been extensively addressed in the field of Privacy-Preserving Data Publishing (PPDP) [15,17] and, more recently, in e-learning systems [18]. A variety of PPDP methods have been proposed to mitigate the risk of re-identifying the data subjects and, in turn, yield protected data that is still useful for specific statistical analyses. Some well-known methods, such as microaggregation [19] or generalization [20], enable data k-anonymization. By k-anonymizing a dataset, each QI is altered—or masked—to make it indistinguishable from the QIs of other $k-1$ individuals, thereby reducing the probability of re-identification to $1/k$ [20]. The k parameter is used to control the masking level: the higher the k, the higher the masking level—and, consequently, the greater the anonymity degree—but the less valuable the anonymized information will be for statistical analysis.

Traditional PPDP techniques that are satisfactory to anonymize single datasets may not be suitable to preserve privacy in the context of data mashup. Privacy-preserving data mashup alternatives not only have to make the integrated dataset satisfy the established privacy requirement, but they also have to face new challenges: (i) the parties involved in the data integration and anonymization process may learn more individuals' information than is disclosed in the integrated and anonymized dataset; (ii) if different attributes from different repositories about the same set of individuals are mashed up—vertically integrated data—a privacy-aware common identifying attribute is required to serve as a link or connector in the integration process; and (iii) mashing up datasets adds new meaning to information that is not available in the individual datasets. These factors may increase the possibility of identifying the individuals' records. PPDP techniques studied in the field of LA [18,21] have not taken into account the distributed nature of learning datasets.

Research Contribution

In this paper, we propose a new *privacy-preserving vertical data (PPVD) mashup protocol* capable of: (i) attending to requests for learning datasets from data consumers; (ii) identifying the learning data sources, i.e., the set of data providers, that can satisfy a particular data request; (iii) vertically integrating learning data from the different educational sources without disclosing the identities of the (students) individuals referenced in the data and *k*-anonymizing the quasi-identifiers of the integrated dataset; and (iv) providing the resulting *k*-anonymized dataset to the data consumer. The protocol integrates learning data effectively and constitutes a Privacy-by-Design (PbD) solution for interoperable data architectures in the educational sphere, reconciling LA with privacy. The protocol can be adopted in any field of application beyond LA systems.

Unlike other privacy-preserving data mashup techniques, our protocol is not linked to a particular *k*-anonymization method. As stated in [14,22], no particular anonymization method is universally the best option for every dataset. Each method has benefits and drawbacks with respect to expected applications of the information. The PPVD mashup protocol offers the possibility of choosing the *k*-anonymization method—either generalization, microaggregation, or any other method that satisfies the *k*-anonymity requirement—according to the nature of the dataset and the utility requirements of the data customers.

We address anonymity in the context of data mashups in terms of unlinkability [23,24] and de-identification [14]. Specifically, we analytically and empirically prove that the proposed protocol is capable of (i) unlinking sensitive data from QIs and (ii) de-identifying sensitive data, thereby preventing adversaries from uniquely associating the sensitive data of a specific (student) individual with their identity.

The PPVD mashup protocol defines a solution to reconcile the seemingly discordant PbD principles [25] and FAIR principles [26]. FAIR principles are essential in the Web of Data and the Knowledge Graph [27], which aim at building wide-scale information systems to share large amounts of data. FAIR principles make it paramount that "metadata clearly and explicitly include the identifier of the data it describes"; "data/metadata include qualified references to other data/metadata"; and "data/metadata are richly described with a plurality of accurate and relevant attributes" [1]. The fulfillment of FAIR principles, however, may threaten data privacy preservation because it requires unveiling identifying attributes and potential quasi-identifier attributes. Since users do not often respect privacy policies [28], enforcing privacy preservation by design is unavoidable for all related works to be considered.

The rest of the paper is organized as follows. Section 2 presents the related works on privacy-preserving data mashups and learning analytics data privacy. Section 3 discusses some underpinning considerations on data mashup required to understand the proposed protocol. Section 4 introduces the PPVD mashup protocol, which is evaluated in Section 5. Finally, in Section 6, the research implications are discussed, and the conclusions are outlined in Section 7.

2. Related Works

Most works on privacy preservation in distributed data environments propose a preemptive approach, defining who has access to private data attributes and resources by defining user profiles [29]. Access control techniques and authentication-based schemes [30] explicitly grant and revoke data access to parties. The larger the number of sensitive attributes in an access-controlled dataset, the greater the loss of analytical utility if exposing only publicly available data attributes. Data partitioning has been proposed as a method for privacy preservation in distributed environments [31]. It is based on a simple PPDP strategy of creating noisy data along with actual data and uploading it to multiple nodes. However, data partitioning approaches are more diverse and have consequences on privacy when applied to data mashups, as we analyze next.

2.1. Privacy-Preserving Data Mashup

Approaches on privacy-preserving data mashups address the integration of *horizontally partitioned* and *vertically partitioned* datasets, each data partition being held by a different data provider.

In the horizontally partitioned datasets, all the data partitions follow the same data schema, i.e., all the data partitions register the same attributes, but each partition contains records of different individuals. A typical scenario of horizontal partition is one in which the data providers are individuals supplying their data, e.g., demographic, health, and exercise data. Privacy-preserving data mashups on horizontally partitioned datasets have been extensively addressed [32–34]. These strategies have in common the segregation of data in the collection process. In the first phase, the data providers send the quasi-identifiers to the data collector, also known as the mashup coordinator. With this information, the mashup coordinator *k*-anonymizes the set of received quasi-identifiers and distributes the masked quasi-identifiers to the data providers. In the following phase, the confidential attributes are sent to the mashup coordinator along with the masked quasi-identifiers. This segregated collection contributes to anonymizing data because it disassociates confidential attributes from original quasi-identifiers. Unlike previous protocols that output *k*-anonymized data, Chamikara et al. [35] presents a perturbative mashup protocol that provides noisy anonymized data to train distributed machine learning models. Data perturbation is caused by geometric data transformations, randomized expansion noise addition, and data shuffling.

In scenarios of vertically partitioned datasets, data providers record different features on the same set of individuals, i.e., each vertical partition registers a different set of attributes and, thus, follows a different data schema. It is assumed that all the vertical partitions have a common identifier attribute, which will be used as a connector to integrate the partitions. Vertical partitioning of data is also a data distribution model often found in real cases, such as healthcare [36], the financial sector [37], or one-stop services [38]. Vertical partitioning is the typical configuration of the datasets used to build the next-generation VLEs. Databases used to store and query e-learning data can be implemented with diverse storage techniques, including graph databases [39], e.g., RDF (Resource Description Framework) triplestores, and relational databases [40].

On the one hand, vertical partitioning in relational database tables to store different sets of data properties [41] is subject to data linkage attacks as long as tables must be subject to relational joins to implement the mashup queries. On the other hand, querying data stored as vertically partitioned graph-based databases [42] can be serialized onto large tables and exposed to privacy concerns. The mashup of data sources in the Web of Data does not only affect distinct data sources from independent RDF triplestores containing resources that can be linked. Even a single-triplestore implementation of the Linked Data Platform (LDP) specification [43] might also require mashing up resources from different containers before running a query because some LDP sources vertically partition their resources into smaller containers, such that each resource is created within an instance of one of these container-like entities. Although containers are not normative in the LDP 1.0 specification, container-based implementations can also be exposed to adversarial attacks affecting data linking from vertically partitioned containers.

Privacy-preserving data mashups on vertically partitioned datasets have been heavily focused on data mining, such as association rule mining [36,44,45], classification mining [46–48], or clustering [49–51]. For example, ref. [36] collaboratively computes association rule mining on vertically partitioned data to find common patterns. The authors of [51] detect the clusters on the integrated dataset by using mechanisms of secure multiparty computation to model a clustering tree on vertically distributed data without revealing the data partitions to other providers or the mashup coordinator. Unlike data integration focused on data mining, data publishing methods are used to share datasets—i.e., raw data—instead of just data mining results—e.g., answers to queries. In many applications, sharing datasets is preferable for flexibility. It allows the data consumers to conduct their

own analysis and data exploration without being linked to any particular query submission protocol. In this regard, Ref. [37] proposes a top-down specialization approach to build the *k*-anonymous datasets from vertical data partitions. The integrated *k*-anonymous dataset is collaboratively built by the data providers from a top-level abstract representation of the dataset. This initial version of the dataset is then specialized down in a sequence of iterations. At each iteration, the provider selected to specialize its quasi-identifiers instructs the other data providers on how to modify those data in the generalized version they keep. The process ends when any further specialization leads to a violation of the *k*-anonymity requirement. Aware of the high dimensionality that the quasi-identifier resulting from the join may have in a vertical data mashup, Ref. [52] proposes a variant to better preserve the information utility on high-dimensional quasi-identifiers.

The techniques on vertically partitioned datasets described above achieve *k*-anonymity by generalizing the dataset, as shown in Table 1. Generalization techniques have the disadvantage of either requiring a high computational cost to find the optimal generalization that minimizes the information loss [53] or requiring an ad hoc taxonomic binary tree for each attribute to be anonymized [54]. It would be desirable to incorporate more practical techniques of *k*-anonymization in vertical data mashups, such as those based on microaggregation [19].

Table 1. Comparison of privacy-preserving data mashup protocols that yield *k*-anonymized raw data.

Privacy-Preserving Data Mashup Protocol	Data Partitioning	Method for *k*-Anonymizing Quasi-Identifiers
Soria-Comas and Domingo-Ferrer (2015) [32]	Horizontal	Any method (e.g., generalization or microaggregation)
Kim and Chung (2019) [33]	Horizontal	Generalization, although other methods can be easily incorporated
Rodriguez-Garcia, Cifredo-Chacón and Quirós-Olozábal (2020) [34]	Horizontal	Any method (e.g., generalization or microaggregation)
Mohammed, Fung, Wang and Hung (2009) [37]	Vertical	Generalization (top-down specialization)
Fung, Trojer, Hung, Xiong, Al-Hussaeni and Dssouli (2012) [52]	Vertical	Generalization (top-down specialization)

2.2. Learning Analytics Data Privacy

Privacy presents severe challenges for current developments and research in the field of ubiquitous [55] and multimodal [56] learning environments. The way students' assignment data are represented in such VLEs is influential to the performance of LA methods and algorithms [57]. For instance, extensions of supervised learning focused on weakly labeled data have been used to predict the impact of students' assignments on their learning [58]. One of the main objectives of the FAIR principles is to enhance these weakly labeled data by enriching metadata with a plurality of attributes. Nevertheless, intelligent computing techniques, such as machine learning, have many security and ethical implications [59] that can be discordant with fulfilling such principles. Consequently, when applied to the arena of technology-enhanced learning, FAIR principles may pose an advantage to humans' learning support as well as a risk on their privacy.

The application of PbD techniques is paramount for LA and analytics research in educational institutions [60]. LA systems development should account for privacy at the time of design rather than addressing privacy concerns as an afterthought [61]. For instance, de-identification helps protect privacy by preventing the revelation of Personal Identifiable Information (PII) that can be used to identify an individual [21]. Besides, quasi-identifiers can also be used to break basic anonymization techniques used for LA [18]. Since current VLEs are built on data from cloud-based environments [7,11], LA requires improved PPDP methods capable of operating on data mashups, such that privacy constraints do not impose a limitation on LA solutions [13]. PPDP solutions used for LA [18,21] have not considered the actual mashup structure of current VLEs.

3. Considerations on Data Mashup

Before describing our PPVD mashup protocol, we have to argue about the database technology options to implement the vertical data mashups based on the most prominent DBMS alternatives for data providers. Then, we present some considerations on each participant's role in mashups of vertically partitioned datasets. Finally, we address sensitive data de-identification in the context of data mashups.

3.1. Database Management Technology

Building a data mashup greatly depends on the DataBase Management System (DBMS) technology data providers use. On the one hand, relational DBMSs are still prevailing to implement internet information systems (72.8% score according to *DB-Engines*' ranking, available at https://db-engines.com/en/ranking_categories, accessed on 9 September 2021). Hence, basing our data mashup implementation upon relational databases would have been reasonable. However, the popularity of graph databases is constantly increasing [39]—graph DBMS are 14 times more prevalent than relational DBMSs. On the other hand, opting for general-purpose graph database technology would require revised versions of PPDP algorithms, which is out of our scope.

An option to share the data mashup schema is using an LDP-compliant triplestore. Despite its currently low popularity (0.4% ranking score), RDF triplestores are graph databases that enable sharing schemata through ontologies and vocabularies. Besides, RDF triplestores are usually implemented on top of relational DBMS or graph databases [40], making LDP an acceptably interoperable solution to publish and share different providers' data schemata, either relational or graph-based. Another choice would have been to use the GraphQL Schema Definition Language (SDL) (available at https://graphql.org/learn/schema, accessed on 9 September 2021) to define both the providers' schema and the data mashup schema. Using Linked Data or GraphQL is an implementation decision that does not affect the validity of the mashup protocol proposed below.

As exchanging the data providers' schemata depends on their implementation choice of DBMS technology, we need an independent means to share the information required by the mashup protocol. We opt for using the Web of Data standards for representing the data mashup, while data providers' vertical partitions are represented as relational tables, as explained next. The choice of relational DMBS as the source of data is supported by the scarce availability of VLE datasets as another format than relational database dumps.

To illustrate the mashup protocol in a real e-learning mashup example, we will use the Open University Learning Analytics Dataset (OULAD) [62], formed by several relational data tables, each concerning different aspects of students' activity in a LMS. Although OULAD is actually a monolithic data dump, it is structured as three parts: (i) student demographics, which contains demographic information about the students together with their results; (ii) student activities, which contains the results of students' assessments and information about the time when the student registered in modules; and (iii) course module presentations, which contains information about available course modules, assessments, and materials. Each OULAD part is considered to be stored by a separate data provider as a more realistic cloud-based, personal learning environment [7] that might involve two or more educational institutions as data providers [11].

3.2. Mashups of Vertically Partitioned Datasets

We consider three actors in a privacy-preserving data mashup protocol: the data consumer, the data provider, and the data mashup coordinator. The *data consumer* is the party that acquires individuals' data for a specific purpose. A data consumer could be, for example, an institution that seeks to acquire LA datasets for sociological studies. The *data provider* is the party that supplies the individuals' data, e.g., an educational center that provides the students' data. Finally, the *data mashup coordinator* represents a point of connection between data consumers and providers. Its function is to coordinate the data

collection, integration, and anonymization. The mashup coordinator may be a third party or the data consumer itself.

The eventual dataset provided to the data consumer is built by vertically joining the data partitions held by a set of providers. Each data partition registers different characteristics—or attributes—from the same set of individuals, and all the partitions share a common key field—or a common identifier attribute. For simplicity, we consider that a *vertical data partition* is a data table. Each row is a data record containing information about a single individual, and each column is an attribute containing information regarding one of the features collected. The attributes of a partition can be classified as identifiers, quasi-identifiers, and confidential. We assume that identifier attributes are not shared with data consumers, except for the common identifier attribute that must be shared with the mashup coordinator in a privacy-preserving manner. We also assume that each data provider decides the amount of masking required for its quasi-identifier attributes.

3.3. De-Identification in the Context of Data Mashup

We consider that data mashup protocols are executed in scenarios where all parties participating in the protocol are semi-honest. A party is semi-honest if, despite following the rules of the protocol, it may attempt to infer additional information—e.g., sensitive information—about the data subjects by analyzing the data received during the execution of the protocol. In this context, we define the k-unlinkability property as a critical property to de-identify sensitive data in data mashups.

Definition 1 (k-Unlinkability)**.** *A data mashup protocol is said to satisfy k-unlinkability if, for any passive attacker, whether internal or external to the protocol, the probability of correctly linking the confidential attributes of a specific individual with their original quasi-identifiers—non-masked quasi-identifiers—is at most* $1/k$.

If this property is not met in a data mashup protocol, an adversary could re-identify sensitive information in the face of successful data linkage attacks.

4. Privacy-Preserving Vertical Data Mashup Protocol

Our proposal consists of two protocols: the *setup protocol* and the *anonymization and integration protocol*. In the setup protocol, the mashup coordinator identifies those data providers that may supply the data partitions used to build the datasets requested by the data consumers. In the anonymization and integration protocol, the data providers and the mashup coordinator k-anonymize and vertically integrate the data partitions to build the de-identified datasets provided to the data consumers.

4.1. Setup Protocol

When the mashup coordinator receives a data request from a particular data consumer, it starts the setup protocol. In the setup protocol, the mashup coordinator must complete the following steps:

1. Identify the set of data providers that can satisfy the data request, each provider contributing a vertical partition of the requested dataset;
2. Build the mashup data schema;
3. Designate the leading provider that will initiate the anonymization and integration protocol.

4.1.1. Identify Data Providers

To facilitate the identification of the data partitions that may be vertically integrated to satisfy the requirement of the data consumer, the data providers must: publish the data schemata, set the identifier attribute that may be used as a connector in the integration process, and define the quasi-identifying and confidential attribute sets. Since publishing the data schemata heavily depends on the DBMS technology used by each data provider,

we suggest mapping to well-known technologies used for the Web of Data to publish the schemata of the involved datasets, as explained in the following subsection.

4.1.2. Build the Mashup Data Schema

Once the mashup coordinator has identified the data providers that can satisfy the data request, the mashup coordinator proceeds to build the final schema of the mashup dataset. This data schema should reflect the following: the identifier attribute that will be used as a connector in the integration process, the *aggregate* (or join) *quasi-identifier*, i.e., the one resulting from joining the quasi-identifiers of the different data partitions, the privacy level that will be applied to the aggregate quasi-identifier, and the set of confidential attributes.

Considering the problem of a large dimension in aggregate quasi-identifiers, we propose that these be divided into smaller quasi-identifiers [52], thereby allowing mashup coordinators to specify multiple aggregate quasi-identifiers. The division of quasi-identifiers prevents a significant loss of information during the masking process because as the number of attributes decreases, less perturbation may be required to achieve k-anonymity. Without loss of generality and for the sake of simplicity, we will describe the protocol for a single aggregate quasi-identifier. To fulfill most exigent data providers' privacy requirements, the mashup coordinator must select the most restrictive k value—i.e., the highest k—from those specified in the data schemata of the providers to be applied to the aggregate quasi-identifier.

In the following, we draw the suggested enactment of the setup protocol in a federated RDF view of an underlying relational data source. We use the RDF view materialization strategy described in [63] to build the linked data mashup. Although it was initially proposed to improve query performance and data availability, we applied it to implement the setup protocol on relational data sources. The mashup materialization depends on the federated schemata, the aggregate schema, the connector attribute, and the quasi-identifiers, as the setup protocol requires.

Concerning the LDP 1.0 specification, two alternative implementations must be considered for the mashup. On the one hand, we can consider a single LDP instance consisting of several resource containers. On the other hand, we can consider several separate LDP instances, each managing their own triplestores. We may restrict the explanation of the implementation to the latter option without losing generality. Besides, it can be a more realistic privacy-preserving scenario, where semi-honest agents from independent LDP instances can be involved.

The materialization of an RDF view is illustrated on OULAD [62]. As for illustrative purposes, we are limiting the description to mashup two data providers: the OULAD Student Demographics (A) and Student Activities (B) parts. We also define the `oula` namespace to map the linked data attributes of the OULAD schema, e.g., `student` or `course`, as long as convenient linked data vocabularies, e.g., foaf and schema.org, might not be easily found or mapped to OULAD data attributes.

- Each tuple t in $A.studentInfo$ produces the following set of RDF triples:

  `oula:student#`$t.id\_student$ `rdf:type foaf:Person`

- For each tuple t in $A.studentInfo$ and each local QI attribute identifiable as such in A, generate one RDF tuple depending on whether there is a corresponding term in a standard linked data vocabulary to map the local QI attribute, as explained next with the OULAD example.
  - If local QI attributes of the $A.studentInfo$ table are considered to be *code_module* and *region*, then generate the following RDF triples—note that, in the case of *code_module* and *region* attributes, no standard vocabulary terms are found or provided:

    `oula:student#`$t.id\_student$ `oula:registeredIn oula:course#`$t.code\_module$
    `oula:student#`$t.id\_student$ `oula:region` $t.studentRegion$

Although we used `oula:region` for *region* instead of a mapping to standard linked data vocabulary terms, a different alignment strategy could determine, for instance, `foaf:based_near` as a valid mapping instead of directly using `oula:region`. Then another option for *region* is to generate an RDF triple, as in the following:

`foaf:based_near owl:sameAs oula:region`

– If for the $A.studentInfo$ schema, students' *gender* is considered as local QI attributes, each tuple t in $A.studentInfo$ would produce one RDF triple —note that, in the case of *gender* attribute, the `schema:gender` term of the *schema.org* vocabulary is selected to map the attribute:

`oula:student#`$t.id\_student$ `schema:gender oula:student#`$t.gender$

As in the previous case, other strategies for vocabulary alignment between the OULAD schema and standard vocabularies can be followed in the case of `oula:gender` values using `oula:gender` instead of `schema:gender` and adding a `owl:sameAs` triple to the generated RDF mashup:

`schema:gender owl:sameAs oula:gender`

The same strategy can be applied to materialize an RDF view that mashes up $A.studentInfo$ and $B.studentAssessment$.

- For each tuple t in $A.studentInfo$ and t' in $B.studentRegistration$ such that $t.id\_student = t'.id\_student$, a triple of the following form is generated for each local QI attribute (e.g., if dates are considered as QI):

`oula:student#`$t.id\_student$ `oula:registeredIn oula:course#`$t'.code\_module$
`oula:student#`$t.id\_student$ `oula:registrationDate oula:course#`$t'.date\_registration$
`oula:student#`$t.id\_student$ `oula:unregistrationDate`
`oula:course#`$t'.date\_unregistration$

- For each tuple t in $A.studentInfo$ and t' in $B.studentAssessment$ such that $t.id\_student = t'.id\_student$, a set of triples of the following form is generated:

`oula:student#`$t.id\_student$ `oula:assessedIn oula:course#`$t'.id\_assessment$
`oula:course#`$t'.id\_assessment$ `oula:submittedBy oula:course#`$t'.date\_submission$
`oula:course#`$t'.id\_assessment$ `oula:scored oula:course#`$t'.score$

Following the setup protocol as applied to the OULAD example, the connector attribute selected is `id_student`, and all the potential QI attributes are the following:

- From `A.studentInfo`:
  - `code_module`
  - `code_presentation`
  - `gender`
  - `region`
  - `highest_education`
  - `imd_band`
  - `age_band`
  - `num_of_prev_attempts`
  - `studied_credits`
  - `disability`
  - `final_result`
  - `date_registration`
  - `date_unregistration`
- From `B.studentAssessment`:
  - `id_assessment`
  - `date_submitted`
  - `is_banked, score`

Thus, we can obtain a mashed-up dataset formed by all or part of the attributes of the previous list. In the mashed-up dataset, an aggregate QI can be determined by any combination of `gender`, `region`, ..., `date_registration`, `date_unregistration`, while `disability`, `final_result`, and `score` are the confidential attributes to be privacy-preserved.

4.1.3. Designate a Leading Provider

Eventually, the mashup coordinator has to connect with the selected providers, inform them about the leading provider, and communicate the schema of the intended dataset. Finally, the coordinator transfers control to the leading provider, which will initiate the anonymization and integration protocol described below. The leading provider can be set by executing a leader election algorithm [64].

4.2. Anonymization and Integration Protocol

This protocol vertically integrates the data partitions identified in the setup protocol and *k*-anonymizes the aggregate quasi-identifier, built by vertically joining the quasi-identifier attributes of each partition. The collection and integration of the vertical partitions of the dataset are carried out by the mashup coordinator and the set of providers participating in the protocol without revealing the individuals' identities in the data. This privacy-preserving data collection and integration is achieved by segregating the quasi-identifier collection from the confidential data collection and by using what we call *privacy-preserving connectors*.

Definition 2 (Privacy-Preserving Connector)**.** *A privacy-preserving connector for a record i of a vertically partitioned dataset, denoted by ppc_i, is a pseudonym of the identifier attribute shared by all the vertical partitions, which is computed as a collision-resistant hash function of the value that the identifier attribute holds in the record i, ID_i, and a nonce common to all records.*

$$ppc_i = H(ID_i, nonce), \qquad 1 \leq i \leq n \tag{1}$$

where n is the number of records in the dataset. The nonce—one-time arbitrary number—is used to prevent reusing the *ppc* and strengthen the *ppc* against dictionary attacks.

The anonymization and integration protocol is detailed below. Figure 1 shows the data transfer among the parties participating in the protocol, and Table 2 lists the symbols and mathematical notations used in the definition of the protocol. Without loss of generality and for the sake of simplicity, we depict the protocol for two data providers, P_a and P_b, each holding a vertical data partition of the final dataset. Each partition contains different quasi-identifier attribute sets—Q^a and Q^b—and different confidential attribute sets—C^a and C^b.

The leading provider selected in the setup protocol (P_a in Figure 1) initiates the anonymization and integration protocol, generating the nonces used to build the privacy-preserving connectors. Two connectors *ppc* are used in the protocol: one to integrate the data partitions received in the quasi-identifier collection, named $Qppc$, and another to integrate the data partitions received in the confidential data collection, named $Cppc$. This segregated collection contributes to anonymizing data because it allows confidential attributes to be disassociated from quasi-identifiers and, thus, prevents the mashup coordinator from linking the original values of the quasi-identifiers with sensitive information. Therefore, the leading provider must generate two nonces: one for each *ppc* (step 1 in Figure 1). These nonces, named *Qnonce* and *Cnonce*, are shared in step 2 with the other data providers participating in the process (P_b in Figure 1) through a secure channel that provides authentication, privacy, and data integrity between communicating parties, such as TLS (Transport Layer Security).

In the quasi-identifier collection, the data providers send their quasi-identifier attributes, Q_i, along with the connector $Qppc$ of each record, $Qppc_i$, ordered by $Qppc_i$, to the mashup coordinator through a secure channel. These data partitions are sent in step 4 of the protocol, the partition of the provider P_a being represented by $(Qppc_i, Q_i^a)_{i=1}^n$, similarly,

for P_b. Previously, as specified in step 3, $Qppc_i$ is derived from $Qnonce$ and ID_i using Equation (1).

Table 2. List of symbols and mathematical notations used in the anonymization and integration protocol.

P_a	data provider a, similarly for the data provider b
ppc	Privacy-Preserving Connector
$Qppc$	ppc used to integrate the data partitions received in the quasi-identifier collection
$Cppc$	ppc used to integrate the data partitions received in the confidential data collection
$Qnonce$	nonce used in the calculation of $Qppc$
$Cnonce$	nonce used in the calculation of $Cppc$
$Qppc_i$	$Qppc$ corresponding to the record i, similarly for $Cppc_i$
$H(.)$	hash function
$(.)_{i=1}^{n}$	set of n records
ID_i	identifier attribute of the record i (held by both P_a and P_b)
Q_i^a	(non-masked) quasi-identifier attributes of the record i held by P_a, similarly for Q_i^b
Q_i^{a*}	masked quasi-identifier attributes of the record i held by P_a, similarly for Q_i^{b*}
C_i^a	confidential attributes of the record i held by P_a, similarly for C_i^b
Q^a	(non-masked) quasi-identifier attributes of the n records held by P_a, similarly for Q^b
Q_{join}	(non-masked) aggregate quasi-identifiers of the n records
Q_{join}^*	masked aggregate quasi-identifiers of the n records

The mashup coordinator vertically integrates the data partitions received in the quasi-identifier collection through the connector $Qppc$, thus building the aggregate quasi-identifier, $Q_{join} = (Qppc_i, Q_i^a, Q_i^b)_{i=1}^n$, as shown in step 5. Then, the mashup coordinator initiates the anonymization process of Q_{join} in step 6. Any PPDP method that satisfies k-anonymity, such as those based on aggregation or generalization mentioned in Section 1, can be used to anonymize the quasi-identifier attributes. The result of the de-identification process is represented by $Q_{join}^* = (Qppc_i, Q_i^{a*}, Q_i^{b*})_{i=1}^n$, Q_i^{a*} and Q_i^{b*} being the masked values of the quasi-identifier attributes of the record i. In step 7, the mashup coordinator sends the anonymized aggregate quasi-identifier set Q_{join}^* to each data provider. Because the anonymization of the quasi-identifiers has been delegated to the mashup coordinator, the data providers must make sure before reporting confidential information that the result satisfies the requirements of k-anonymity. Each provider must check that the k-anonymous groups in Q_{join}^* comprise k or more records.

Once Q_{join}^* is received, each data provider integrates Q_{join}^* with the confidential data of its data partition to form the confidential data partition (step 8). This integration is achieved through the connector $Qppc$. Then, in step 9, each data provider sends its confidential data partition along with the connectors $Cppc_i$, ordered by $Cppc_i$, to the mashup coordinator through a secure channel—e.g., the data set sent by the provider P_a is $(Cppc_i, Q_i^{a*}, Q_i^{b*}, C_i^{a*})_{i=1}^n$, similarly for P_b. Note that the connector $Cppc_i$ of each record is derived from $Cnonce$ and ID_i in step 3. Finally, as shown in step 10, the mashup coordinator vertically joins the received confidential data partitions through the connector $Cppc_i$ to yield the de-identified dataset provided to the data consumer. This dataset, $(Q_i^{a*}, Q_i^{b*}, C_i^a, C_i^b)_{i=1}^n$, satisfies k-anonymity because at least k records share the same values in the aggregate quasi-identifier.

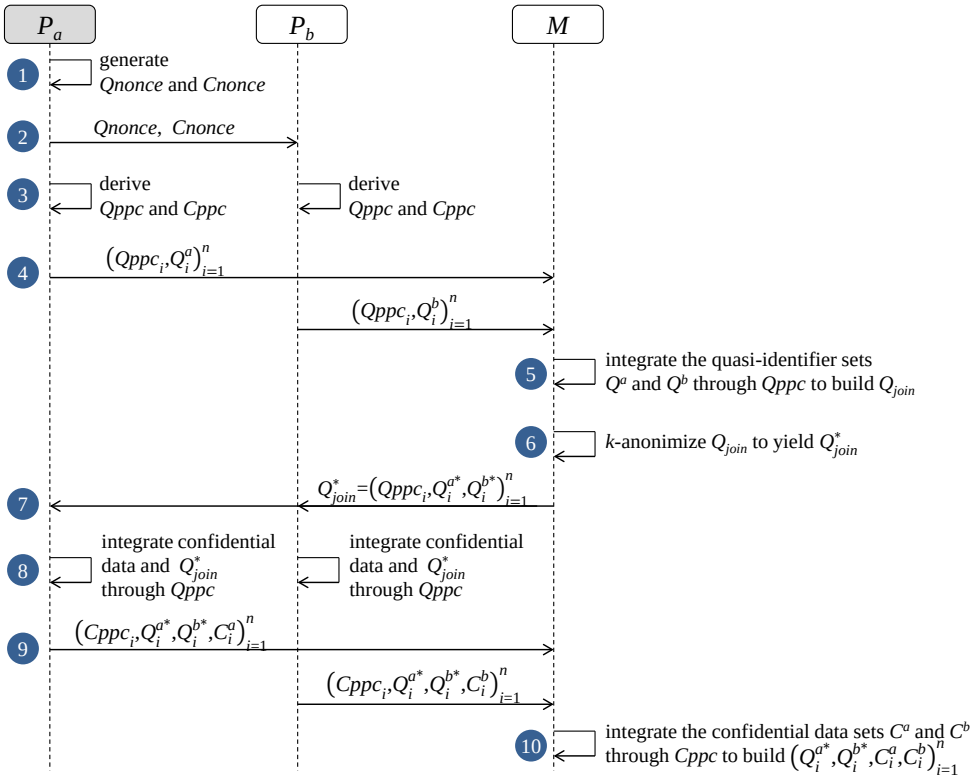


Figure 1. Anonymization and integration protocol in a scenario with two data providers, P_a and P_b, and a mashup coordinator, M. The provider P_a acts as a leading provider.

5. Evaluation

In this section, first, we perform an analytical evaluation of privacy. Specifically, we evaluate whether our protocol can prevent passive adversaries from unambiguously associating the confidential attributes of a particular individual with their original quasi-identifiers (*k*-unlinkability property) and, consequently, from re-identifying their sensitive data. We assume that any participant in the anonymization and integration protocol, whether a data provider or the mashup coordinator, is a potential adversary and may be interested in inferring information about the data subjects. Secondly, we empirically evaluate whether the proposed protocol achieves the *k*-unlinkability, and consequently, the sensitive data can no longer be identified.

5.1. Analytical Evaluation of the k-Unlinkability Property

We evaluate whether our protocol satisfies the *k*-unlinkability property. The evaluation is conducted in the worst-case scenario when the passive adversary participates in the anonymization and integration protocol.

When the passive adversary is a data provider participating in the protocol:

(i) Because the data partitions are sent encrypted to the mashup coordinator through a secure transport protocol, no data provider will be able to view other providers' quasi-identifier and confidential attributes, even if the provider carried out a network traffic analysis.

Based on (i), we conclude that a malicious provider cannot associate the confidential data of a particular individual with their original quasi-identifiers because those data are unknown to the provider.

When the passive adversary is the mashup coordinator participating in the protocol:

(ii) Because the mashup coordinator handles quasi-identifiers and confidential attributes during the execution of the protocol, the coordinator may learn additional information about the subjects of those data by linking the data obtained in the different steps of

the protocol. After analyzing the data handled by the mashup coordinator, compiled in Table 3, it follows that the mashup coordinator can only make ambiguous links between confidential attributes and original quasi-identifiers. In particular, the mashup coordinator can only perform the following reverse linking of the information:

$$(C_i^a, C_i^b) \longrightarrow (Q_i^{a*}, Q_i^{b*}) \longrightarrow (Qppc_i)_{i=1}^k \longrightarrow (Q_i^a, Q_i^b)_{i=1}^k$$

That is, in step 10 of the protocol, the mashup coordinator can link the confidential attributes (C_i^a, C_i^b) with the masked quasi-identifier attributes (Q_i^{a*}, Q_i^{b*}). In turn, (Q_i^{a*}, Q_i^{b*}) can be linked with k or more connectors $Qppc$ by using the k-anonymized data from step 6. Note that the masked quasi-identifiers of a given individual can never be linked with less than k connectors since, after k-anonymization, the number of privacy-preserving connectors that have associated the same values in the masked quasi-identifier attributes is always greater than or equal to k. Finally, from the data received in step 4, the mashup coordinator can link the k (or more) connectors $Qppc$ with their respective original quasi-identifiers. Because the connectors used in step 10, $Cppc$, are different from those received in step 4, $Qppc$, the mashup coordinator will never be able to link a given $Cppc_i$ with its corresponding $Qppc_i$, and thus, it will not be able to uniquely associate the confidential attributes from a given individual with their original quasi-identifiers.

Table 3. Data handled by the mashup coordinator during the execution of the anonymization and integration protocol.

Protocol Step	Receive	Integrate	k-Anonymize
Step 4	$(Qppc_i, Q_i^a)_{i=1}^n$ $(Qppc_i, Q_i^b)_{i=1}^n$		
Step 5		$(Qppc_i, Q_i^a, Q_i^b)_{i=1}^n$	
Step 6			$(Qppc_i, Q_i^{a*}, Q_i^{b*})_{i=1}^n$
Step 9	$(Cppc_i, Q_i^{a*}, Q_i^{b*}, C_i^a)_{i=1}^n$ $(Cppc_i, Q_i^{a*}, Q_i^{b*}, C_i^b)_{i=1}^n$		
Step 10		$(Cppc_i, Q_i^{a*}, Q_i^{b*}, C_i^a, C_i^b)_{i=1}^n$	

Based on (ii), we conclude that a malicious mashup coordinator can at most associate the confidential data of a particular individual with the set of original quasi-identifiers of the k-anonymous group to which the individual belongs. Therefore, the probability that the mashup coordinator correctly correlates the confidential attributes to the original quasi-identifiers is at most $1/k$. The higher the value of k, the greater the uncertainty of the mashup coordinator.

Therefore, the proposed anonymization and integration protocol satisfies the k-unlinkability property, whether the passive adversary is a data provider or the mashup coordinator.

5.2. Analytical Evaluation of the De-Identification of Sensitive Data

We evaluate whether our protocol is capable of de-identifying the sensitive data collected during the anonymization and integration process. We evaluate this feature by analyzing the probability of re-identification of the sensitive data collected. The evaluation is conducted in the worst-case scenario, that is, when the passive adversary is the mashup coordinator since it follows from Section 5.1 that the mashup coordinator is the only party that knows the original values (non-masked) of the aggregate quasi-identifier.

When the passive adversary is the mashup coordinator participating in the protocol:

(i) Because the mashup coordinator handles the original quasi-identifiers during the execution of the protocol, the mashup coordinator may associate them with the connectors $Qpcc$.

(ii) Because a connector $Qpcc$ results from a one-way hash function on a nonce and the individual's identifier attribute (both unknown to the mashup coordinator), the

mashup coordinator will not be able to derive the value of the identifier. Moreover, if the nonce is large enough, the connector will be protected against dictionary attacks and other precomputation attacks, making such attacks infeasible.

Based on (i) and (ii), we conclude that a malicious mashup coordinator cannot associate the original quasi-identifier attributes of a given individual with their identifier attribute, even if the mashup coordinator carried out a dictionary attack, or similar. Despite not being able to re-identify a record using information learned from the protocol, the mashup coordinator could attempt re-identification through data linkage attacks or re-identification attacks [15]. If re-identification was successful, the mashup coordinator could not unambiguously link the individual's identity with their sensitive information because the proposed protocol satisfies the k-unlinkability property, being the probability of success of this link less than or equal to $1/k$.

Therefore, the proposed anonymization and integration protocol is capable of de-identifying the sensitive data collected, such that the probability that the mashup coordinator re-identifies the sensitive data of an individual is at most $1/k$.

5.3. Empirical Evaluation

This subsection empirically evaluates whether the proposed PPVD mashup protocol achieves k-unlinkability between the quasi-identifier and confidential attributes and, consequently, de-identifying sensitive data against passive adversaries.

We used a simulated scenario with two data providers and one mashup coordinator to conduct the empirical evaluation. The data providers hold data partitions of OULAD [62], which contains data about courses, students, and their interactions with a VLE. Specifically, the provider P_a holds the *OULAD studentInfo* as a vertical data partition with diverse demographic information about 342 students, plus their final results in the courses; the provider P_b holds the *OULAD studentAssessment* as another vertical data partition, containing the results of a specific learning assessment ($id\_assessment = 1753$). The attribute used as a connector in the execution of the protocol is the identifier attribute *id_student*. The attributes to be vertically joined are indicated in Table 4. We had to adjust the *age* attribute in the data partition of P_A since OULAD already provides masked values for this attribute. In particular, the values of the *age* attribute are generalized in three ranges: 0–35, 33–55, and >55. Since our protocol operates on non-anonymized data, for each *id_student*, a random synthetic value between the two limits was assigned. Data records with *age* in the range 0–35 were assigned a value between 18 and 35; those between 35 and 55 were assigned a value between 36 and 55; and those greater than 55 received a value between 56 and 75.

Table 4. Attributes of the data partitions held by P_a and P_b.

Data Partition	Attribute	Type	Description
$A.studentInfo$	age	quasi-identifier	age of the student
$A.studentInfo$	disability	confidential	indicates whether the student has declared a disability
$A.studentInfo$	final_result	confidential	student's final result
$B.studentAssessment$	date_submitted	quasi-identifier	date of student submission, measured as the number of days since the start of the module presentation
$B.studentAssessment$	score	confidential	student's score in this assessment

The protocol was evaluated in the worst-case scenario—when the passive adversary is the mashup coordinator. This party handles original quasi-identifiers and confidential attributes during the execution of the protocol, and, as discussed in Section 5.1, it may have more information than any other adversary. The method used to k-anonymize the set of aggregate quasi-identifiers resulting from the data mashup was the multivariate microaggregation method (using the mean as an aggregate) with a privacy parameter k equal to 5. The privacy-preserving connectors, $Qppc$ and $Cppc$, were built on the attribute *id_student* using nonces of 128 bits and the SHA-256 as Cryptographic Hash function.

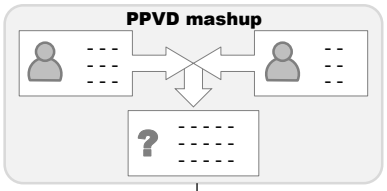
Figure 2 illustrates the result of the execution of the PPVD mashup protocol on an excerpt of 10 records. The web address with the full versions of the vertical data partitions and the output dataset is published in the Data Availability Statement section.

Student ID	**Age**	Disability	Final result
141355	**29**	N	Pass
2411778	**33**	N	Withdrawn
236284	**29**	N	Pass
205719	**22**	Y	Fail
2376496	**29**	N	Pass
11391	**65**	N	Pass
186149	**38**	N	Pass
1401935	**51**	N	Fail
2536991	**64**	N	Distinction
106247	**46**	N	Withdrawn

Vertical partition of the provider P_a

Student ID	**Submission**	Score
141355	**54**	74
2411778	**52**	83
236284	**54**	84
205719	**58**	63
2376496	**54**	72
11391	**53**	85
186149	**68**	81
1401935	**61**	25
2536991	**54**	72
106247	**64**	66

Vertical partition of the provider P_b



Age*	**Submission***	Score	Disability	Final result
28	**54**	74	N	Pass
28	**54**	83	N	Withdrawn
28	**54**	84	N	Pass
28	**54**	63	Y	Fail
28	**54**	72	N	Pass
53	**60**	85	N	Pass
53	**60**	81	N	Pass
53	**60**	25	N	Fail
53	**60**	72	N	Distinction
53	**60**	66	N	Withdrawn

Integrated and *k*-anonymized dataset

Figure 2. Execution result of the PPVD mashup protocol on an excerpt of 10 records. Quasi-identifier attributes are marked in bold. The asterisk identifies the masked attributes.

To verify whether the proposed protocol fulfills *k*-unlinkability and, consequently, is capable of de-identifying sensitive data, we analyzed the information that the mashup coordinator handled during the execution of the protocol. In the quasi-identifier collection carried out in step 4 of the protocol, the mashup coordinator obtained the original quasi-identifiers of the 342 students along with the privacy-preserving connectors $Qppc$, ordered by $Qppc$. An extract of 10 records is shown in Figure 3. We have changed the order of the records to clarify the illustration. Because $Qppc$ is the result of a one-way hash function strengthened with a nonce, the mashup coordinator could not derive the students' identifiers in a feasible computational time and, thus, re-identify the records. After integrating the quasi-identifier attributes of each student to form the aggregate quasi-identifier $Q_{join} = (age, submission)$, the mashup coordinator *k*-anonymized the aggregate quasi-identifier values of the 342 students with $k = 5$. As expected, the anonymization process resulted in a dataset consisting of *k*-anonymous groups (5-anonymous groups), with the number of records in each *k*-anonymous group always greater than or equal to 5. Each generated 5-anonymous group contains the masked aggregate quasi-identifier for that group and the connectors $Qppc$ of the students belonging to the group. As shown in

Figure 3, the 5-anonymous groups are formed by 5 records of students, i.e., 5 $Qppc$, and the masked aggregate quasi-identifier of the group, e.g., the last 5-anonymous group has the masked aggregate quasi-identifier (*age**, *submission**) = (53, 60).

(a)

Qppc			**Age**
2579a4dddb201431cdfdc91aa2bd1d74 29b1610be4239da3386d8a340419c1d8			**29**
7977	…	3454	**33**
09ee	…	4589	**29**
818a	…	ae60	**22**
4c4c	…	ca7e	**29**
4217	…	1053	**65**
4b74	…	f067	**38**
e353	…	8ba8	**51**
c0e5	…	7948	**64**
bf06	…	a5d8	**46**

(b)

Qppc			**Submission**
2579a4dddb201431cdfdc91aa2bd1d74 29b1610be4239da3386d8a340419c1d8			**54**
7977	…	3454	**52**
09ee	…	4589	**54**
818a	…	ae60	**58**
4c4c	…	ca7e	**54**
4217	…	1053	**53**
4b74	…	f067	**68**
e353	…	8ba8	**61**
c0e5	…	7948	**54**
bf06	…	a5d8	**64**

(c)

Qppc			**Age***	**Submission***	
2579a4dddb201431cdfdc91aa2bd1d74 29b1610be4239da3386d8a340419c1d8			**28**	**54**	
7977	…	3454	**28**	**54**	
09ee	…	4589	**28**	**54**	*k*-anonymous group
818a	…	ae60	**28**	**54**	
4c4c	…	ca7e	**28**	**54**	
4217	…	1053	**53**	**60**	
4b74	…	f067	**53**	**60**	
e353	…	8ba8	**53**	**60**	*k*-anonymous group
c0e5	…	7948	**53**	**60**	
bf06	…	a5d8	**53**	**60**	

Figure 3. (**a**) Data partition sent by the provider P_a during the quasi-identifier collection (step 4 of the protocol). (**b**) Data partition sent by the provider P_b during the quasi-identifier collection (step 4 of the protocol). (**c**) *k*-anonymized and integrated quasi-identifiers with $k = 5$ (steps 5 and 6 of the protocol). Quasi-identifier attributes are marked in bold. The asterisk identifies the masked attributes.

In the confidential data collection carried out in step 9 of the protocol, the mashup coordinator obtained the confidential attributes of the 342 students along with the privacy-preserving connectors $Cppc$ and the masked aggregate quasi-identifiers. Because the connectors in the confidential data collection, $Cppc$, were different from those used in the quasi-identifier collection, $Qppc$, the mashup coordinator could not link $Cppc$ with their corresponding $Qppc$, causing the dissociation between the confidential attributes and the original quasi-identifiers. The mashup coordinator only succeeded in making ambiguous associations. The students' confidential data were associated with the quasi-identifiers of at least five students, those belonging to their 5-anonymous groups. The effects of the 5-unlinkability can be verified on any record of those shown in Figure 4. For example, the confidential attributes $(score, disability, finalresult) = (25, N, fail)$ of the eighth student cannot be linked to their original quasi-identifier attributes $(age, submission) = (51, 61)$ because the connector $Cppc = deb6\ldots e63b$ does not match $Qppc = e353...8ba8$. By using the student's masked aggregate quasi-identifier $(53, 60)$, the confidential attributes $(25, N, fail)$ can be linked to at least five different original aggregate quasi-identifiers $(65, 53)$, $(38, 68)$, $(51, 61)$, $(64, 54)$,

and $(46, 64)$. Therefore, the probability that the mashup coordinator correctly correlates the confidential attributes of the eighth student with their identity is at most $1/5$.

We can conclude that the mashup coordinator's probability of correctly correlating the confidential attributes to the original quasi-identifiers is $1/k$ at most, thus verifying the k-unlinkability property. Logically, the uncertainty will be higher for a higher value of k. As a consequence of the k-unlinkability property, if the mashup coordinator had re-identified the students through data linkage attacks by using their original quasi-identifiers and external data sources, it would not be able to link students' identities to their confidential attributes unambiguously. The experiment, thus, shows that the proposed PPVD mashup protocol is capable of de-identifying the sensitive data collected, such that the probability that an adversary re-identifies the sensitive data of an individual is $1/k$ at most.

(a)

Cppc			**Age***	**Submission***	Disability	Final result
ac64dc171f290c502f4caf5593888b3c ecc3d6f352d6485d9135991b98ef2fb7			**28**	**54**	N	Pass
d91f	…	3b03	**28**	**54**	N	Withdrawn
7eb8	…	222d	**28**	**54**	N	Pass
6d90	…	9ba0	**28**	**54**	Y	Fail
194a	…	ed20	**28**	**54**	N	Pass
fe41	…	d800	**53**	**60**	N	Pass
9584	…	1b3a	**53**	**60**	N	Pass
deb6	…	e63b	**53**	**60**	N	Fail
8041	…	a485	**53**	**60**	N	Distinction
15e7	…	c657	**53**	**60**	N	Withdrawn

(b)

Cppc			**Age***	**Submission***	Score
ac64dc171f290c502f4caf5593888b3c ecc3d6f352d6485d9135991b98ef2fb7			**28**	**54**	74
d91f	…	3b03	**28**	**54**	83
7eb8	…	222d	**28**	**54**	84
6d90	…	9ba0	**28**	**54**	63
194a	…	ed20	**28**	**54**	72
fe41	…	d800	**53**	**60**	85
9584	…	1b3a	**53**	**60**	81
deb6	…	e63b	**53**	**60**	25
8041	…	a485	**53**	**60**	72
15e7	…	c657	**53**	**60**	66

Figure 4. (**a**) Data partition sent by the provider P_a during the confidential data collection (step 9 of the protocol). (**b**) Data partition sent by the provider P_b during the confidential data collection (step 9 of the protocol). Quasi-identifier attributes are marked in bold. The asterisk identifies the masked attributes.

6. Discussion

Internet information systems and applications often use personal information, thus requiring a conservative treatment of PII and confidential information. Our PPVD mashup protocol has implications in the design and construction of information systems on the Internet.

The privacy-by-design challenge is being tackled in the Web of Data by putting individuals in control of their own data through Personal Data Ecosystems based on SOLID principles [65]. In SOLID, individuals store their data on the Web as personal data stores or *pods*, such that each user has one or more pod from different web providers. Applications can access users' data using decentralized authentication and access control mechanisms to guarantee the privacy of the data. Web protocols and access control mechanisms do not sufficiently ensure users' data privacy as long as an adversary can mash up data from several pods and run data linkage attacks.

The Web of Things is also a key driver to understand the paradigm shift in e-learning towards context-aware, ubiquitous learning [66]. Internet-of-Things (IoT) technologies are convenient data gathering systems to build cooperative information systems on the Internet with different purposes, including e-learning. *Things*, such as devices enabled with computational and data storage capabilities, lay the foundations of cloud, fog, and edge computing [67] as the most recent trends in IoT-distributed computing. All such approaches have in common: the system data are spread over multiple devices and must be mashed up in a central point before making computer-aided data-informed decisions. IoT-based information systems, however, also pose a challenge to personal data privacy [68]. The paradigm shift of smart devices connected to the Internet requires considering data mashups in the Web of Things [69]. *Things* are more prone to security risks because digital users' privacy is a fundamental right [70]. Among all security and privacy issues [71], the mashup of sparsely distributed data in the Web of Things is vulnerable to data linkage attacks by semi-honest intermediate entities part of the cloud, fog, or edge computing network infrastructure [72].

Most related works propose preemptive access control [29] and authentication schemes [30] for data mashup privacy preservation in fog computing environments. However, the utility of the published data is lesser in such approaches because they are limited by design to expose publicly available data only. Instead, our protocol can publish all data attributes in a data mashup required for statistical analyses. It makes it with the help of a PPDP method of choice, which is independent of the actual mashup strategy. In contrast, privacy-preserving data partitioning solutions used in fog computing environments are based on simple noise addition [31].

7. Conclusions

In this paper, we have presented a new privacy-preserving data mashup protocol capable of vertically integrating data partitions from multiple educational sources to satisfy the data consumers' requests without disclosing the identities of the individuals referenced in the data. Educational information to be integrated and anonymized typically comes from cloud-based e-learning environments and includes attendance to course activities, course evaluations, feedback on course materials and teaching systems, performance records, and social network data of students and instructors. Our protocol can de-identify the fused information by k-anonymizing the aggregate quasi-identifiers, resulting from the mashup of the data partitions. Unlike other privacy-preserving data mashup techniques on vertically partitioned datasets, our protocol is not linked to a particular k-anonymization method. Therefore, the protocol offers the possibility of choosing the k-anonymization method—either generalization, microaggregation or any method that satisfies the k-anonymity requirement—according to the dataset scheme and to the utility requirements of the data customers.

Our protocol is capable of preventing passive adversaries, whether internal or external to the anonymization and integration process, from re-identifying individuals' sensitive data. In particular, the probability that an adversary correctly correlates the confidential attributes of an individual with their identity is $1/k$ at most. The privacy parameter k thus determines the degree of uncertainty of the adversary. The analytic utility of the protected data is conditioned by the selected method of k-anonymization.

The implementation of the proposed protocol is based on linked data, considering several separate linked data platform instances. A linked data-based implementation provides a shared architecture for linking the information contained in the different educational sources and effectively avoids ambiguity. The use of privatized and shared datasets in the Web of Data compliant with FAIR and privacy-by-design principles enables learning analytics while safeguarding the students' data privacy. With the linked data-based implementation, the datasets involved in the mashups can be semantically described, indicating which are the quasi-identifiers and the sensitive data. Thus, our mashup protocol

enables the combination of datasets increasing privacy-by-design without undermining FAIR principles.

Author Contributions: All authors have contributed to the manuscript according to the following tasks: Conceptualization, M.R.-G. and J.M.D.; methodology, M.R.-G.; validation, M.R.-G., A.B. and J.M.D.; data curation, M.R.-G. and A.B.; writing—original draft preparation, M.R.-G., J.M.D. and A.B.; writing—review and editing, M.R.-G., J.M.D. and A.B.; visualization, M.R.-G.; supervision and project administration, J.M.D.; funding acquisition, J.M.D. All authors have read and agreed to the published version of the manuscript.

Funding: The Spanish National Research Agency (AEI) funded this research through the project CRÊPES (ref. PID2020-115844RB-I00) with ERDF funds.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The OULAD dataset [62] used to support the reported results can be found at https://analyse.kmi.open.ac.uk/open_dataset (accessed on 9 September 2021). According to their curators' description, it contains data about courses, students, and their interactions with a VLE for seven selected courses, called modules. Data illustrating the PPVD mashup protocol execution can be found in https://doi.org/10.5281/zenodo.5411994 (accessed on 3 September 2021).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DBMS	DataBase Management System
FAIR	Findability, Accessibility, Interoperability, and Reusability
FERPA	Family Educational Rights and Privacy Act
IoT	Internet-of-Things
LA	Learning Analytics
LDP	Linked Data Platform
LMS	Learning Management Systems
NIST	National Institute of Standards and Technology
OULAD	Open University Learning Analytics Dataset
PbD	Privacy-by-Design
PII	Personal Identifiable Information
PPDP	Privacy-Preserving Data Publishing
PPVD	Privacy-Preserving Vertical Data
QI	Quasi-Identifiers
RDF	Resource Description Framework
SDL	Schema Definition Language
TLS	Transport Layer Security
VLE	Virtual Learning Environments

References

1. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
2. IEEE Big Data Governance and Metadata Management, Industry Connections Activity. Big Data Governance and Metadata Management: Standards Roadmap. Available online: https://standards.ieee.org/content/dam/ieee-standards/standards/web/governance/iccom/bdgmm-standards-roadmap-2020.pdf (accessed on 9 September 2021).
3. Chang, W.; Mishra, S.; NIST, N.P. *NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2015. [CrossRef]
4. Chang, W.; Boyd, D.; Levin, O. *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2019. [CrossRef]

5. Chang, W.; Reinsch, R.; Boyd, D.; Buffington, C. *NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2019. [CrossRef]
6. Open Data Center Alliance. Big Data Consumer Guide. Available online: https://bigdatawg.nist.gov/_uploadfiles/M0069_v1_7760548891.pdf (accessed on 9 September 2021).
7. Ko, C.C.; Young, S.S.C. Explore the Next Generation of Cloud-Based E-Learning Environment. In Proceedings of the International Conference on Technologies for E-Learning and Digital Entertainment, Taipei, Taiwan, 7–9 September 2011; Chang, M., Hwang, W.Y., Chen, M.P., Müller, W., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6872, pp. 107–114. [CrossRef]
8. Wild, F.; Mödritscher, F.; Sigurdarson, S. Mash-Up Personal Learning Environments. In *E-Infrastructures and Technologies for Lifelong Learning: Next Generation Environments*; Magoulas, G., Ed.; IGI Global: Hershey, PA, USA, 2011; pp. 126–149. [CrossRef]
9. Rodosthenous, C.T.; Kameas, A.D.; Pintelas, P. Diplek: An Open LMS that Supports Fast Composition of Educational Services. In *E-Infrastructures and Technologies for Lifelong Learning: Next Generation Environments*; Magoulas, G., Ed.; IGI Global: Hershey, PA, USA, 2011; pp. 59–89. [CrossRef]
10. Wurzinger, G.; Chang, V.; Guetl, C. Towards greater flexibility in the learning ecosystem—Promises and obstacles of service composition for learning environments. In Proceedings of the 3rd IEEE International Conference on Digital Ecosystems and Technologies, Istanbul, Turkey, 1–3 June 2009; pp. 241–246. [CrossRef]
11. Conde, M.A.; Hernández-García, A. Data Driven Education in Personal Learning Environments—What about Learning beyond the Institution? *Int. J. Learn. Anal. Artif. Intell. Educ.* **2019**, *1*. [CrossRef]
12. Mangaroska, K.; Vesin, B.; Kostakos, V.; Brusilovsky, P.; Giannakos, M.N. Architecting Analytics Across Multiple E-Learning Systems to Enhance Learning Design. *IEEE Trans. Learn. Technol.* **2021**, *14*, 173–188. [CrossRef]
13. Griffiths, D.; Drachsler, H.; Kickmeier-Rust, M.; Steiner, C.; Hoel, T.; Greller, W. Is Privacy a Show-stopper for Learning Analytics? A Review of Current Issues and their Solutions. *Learn. Anal. Rev.* **2016**, *6*, 1–30, ISSN 2057-7494.
14. U.S. Department of Education. Family Educational Rights and Privacy Act, 34 CFR §99 (FERPA). Available online: https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html (accessed on 9 September 2021).
15. Hundepool, A.; Domingo-Ferrer, J.; Franconi, L.; Giessing, S.; Nordholt, E.S.; Spicer, K.; de Wolf, P.P. *Statistical Disclosure Control*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2012.
16. Chang, W.; Roy, A.; Underwood, M. *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2019. [CrossRef]
17. Fung, B.C.M.; Wang, K.; Chen, R.; Yu, P.S. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Comput. Surv.* **2010**, *42*, 1–53. [CrossRef]
18. Gursoy, M.E.; Inan, A.; Nergiz, M.E.; Saygin, Y. Privacy-Preserving Learning Analytics: Challenges and Techniques. *IEEE Trans. Learn. Technol.* **2017**, *10*, 68–81. [CrossRef]
19. Domingo-Ferrer, J.; Torra, V. Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. *Data Min. Knowl. Discov.* **2005**, *11*, 195–212. [CrossRef]
20. Samarati, P. Protecting Respondents' Identities in Microdata Release. *IEEE Trans. Knowl. Data Eng.* **2001**, *13*, 1010–1027. [CrossRef]
21. Khalil, M.; Ebner, M. De-Identification in Learning Analytics. *J. Learn. Anal.* **2016**, *3*, 129–138. [CrossRef]
22. U.S. Office for Civil Rights. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Available online: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (accessed on 9 September 2021).
23. Bleumer, G. Unlinkability. In *Encyclopedia of Cryptography and Security*; van Tilborg, H.C.A., Jajodia, S., Eds.; Springer: Boston, MA, USA, 2011; p. 1350. [CrossRef]
24. Katos, V. Managing IS Security and Privacy. In *Encyclopedia of Information Science and Technology*, 2nd ed.; Khosrow-Pour, M., Ed.; IGI Global: Hershey, PA, USA, 2009; pp. 2497–2503. [CrossRef]
25. Cavoukian, A. Privacy by Design: The 7 Foundational Principles. Available online: https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf (accessed on 9 September 2021).
26. Wilkinson, M.D.; Verborgh, R.; da Silva Santos, L.O.B.; Clark, T.; Swertz, M.A.; Kelpin, F.D.; Gray, A.J.; Schultes, E.A.; van Mulligen, E.M.; Ciccarese, P.; et al. Interoperability and FAIRness through a novel combination of Web technologies. *Peerj Comput. Sci.* **2017**, *3*. [CrossRef]
27. Singhal, A. Introducing the Knowledge Graph: Things, Not Strings. Official Blog of Google, 2012. Available online: http://goo.gl/zivFV (accessed on 9 September 2021).
28. Obar, J.A.; Oeldorf-Hirsch, A. The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. *Inf. Commun. Soc.* **2020**, *23*, 128–147. [CrossRef]
29. Cesconetto, J.; Augusto Silva, L.; Bortoluzzi, F.; Navarro-Cáceres, M.; Zeferino, C.A.; Leithardt, V.R.Q. PRIPRO-Privacy Profiles: User Profiling Management for Smart Environments. *Electronics* **2020**, *9*, 1519. [CrossRef]
30. Patwary, A.A.N.; Fu, A.; Battula, S.K.; Naha, R.K.; Garg, S.; Mahanti, A. FogAuthChain: A secure location-based authentication scheme in fog computing environments using Blockchain. *Comput. Commun.* **2020**, *162*, 212–224. [CrossRef]
31. Patwary, A.A.N.; Naha, R.K.; Garg, S.; Battula, S.K.; Patwary, M.A.K.; Aghasian, E.; Amin, M.B.; Mahanti, A.; Gong, M. Towards Secure Fog Computing: A Survey on Trust Management, Privacy, Authentication, Threats and Access Control. *Electronics* **2021**, *10*, 1171. [CrossRef]

32. Soria-Comas, J.; Domingo-Ferrer, J. Co-utile Collaborative Anonymization of Microdata. In Proceedings of the 12th International Conference on Modeling Decisions for Artificial Intelligence, Skövde, Sweden, 21–23 September 2015; Torra, V., Narukawa, Y., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9321, pp. 192–206. [CrossRef]
33. Kim, S.; Chung, Y. An anonymization protocol for continuous and dynamic privacy-preserving data collection. *Future Gener. Comput. Syst.* **2019**, *93*, 1065–1073. [CrossRef]
34. Rodriguez-Garcia, M.; Cifredo-Chacón, M.A.; Quirós-Olozábal, A. Cooperative Privacy-Preserving Data Collection Protocol Based on Delocalized-Record Chains. *IEEE Access* **2020**, *8*, 180738–180749. [CrossRef]
35. Chamikara, M.; Bertok, P.; Khalil, I.; Liu, D.; Camtepe, S. Privacy preserving distributed machine learning with federated learning. *Comput. Commun.* **2021**, *171*, 112–125. [CrossRef]
36. Domadiya, N.; Rao, U.P. Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data. *Procedia Comput. Sci.* **2019**, *148*, 303–312. [CrossRef]
37. Mohammed, N.; Fung, B.C.M.; Wang, K.; Hung, P.C.K. Privacy-Preserving Data Mashup. In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '09), St. Petersburg, Russia, 23–25 March 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 228–239. [CrossRef]
38. Flumian, M. *The Management of Integrated Service Delivery: Lessons from Canada*; Number 6; Inter-American Development Bank: Washington, DC, USA, 2018. [CrossRef]
39. Sakr, S.; Bonifati, A.; Voigt, H.; Iosup, A.; Ammar, K.; Angles, R.; Aref, W.; Arenas, M.; Besta, M.; Boncz, P.A.; et al. The Future Is Big Graphs: A Community View on Graph Processing Systems. *Commun. ACM* **2021**, *64*, 62–71. [CrossRef]
40. Ali, W.; Yao, B.; Saleem, M.; Hogan, A.; Ngomo, A.C.N. Survey of RDF Stores & SPARQL Engines for Querying Knowledge Graphs. *TechRXiv* **2021**. [CrossRef]
41. Abadi, D.J.; Marcus, A.; Madden, S.; Hollenbach, K. SW-Store: A vertically partitioned DBMS for Semantic Web data management. *J. Very Large Data Bases* **2009**, *18*, 385–406. [CrossRef]
42. Ingalalli, V.; Ienco, D.; Poncelet, P. Chapter 5: Querying RDF Data: A Multigraph-based Approach. In *NoSQL Data Models: Trends and Challenges*; John Wiley & Sons: Hoboken, NJ, USA, 2018; Volume 1, pp. 135–165. [CrossRef]
43. Speicher, S.; Arwe, J.; Malhotra, A. Linked Data Platform 1.0 W3C Recommendation. Available online: https://www.w3.org/TR/ldp/ (accessed on 9 September 2021).
44. Vaidya, J.; Clifton, C. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02), Edmonton, AB, Canada, 23–26 July 2002; Association for Computing Machinery: New York, NY, USA, 2002; pp. 639–644. [CrossRef]
45. Vaidya, J.; Clifton, C. Secure set intersection cardinality with application to association rule mining. *J. Comput. Sci.* **2005**, *13*, 593–622. [CrossRef]
46. Vaidya, J.; Clifton, C. Privacy Preserving Naive Bayes Classifier for Vertically Partitioned Data. In Proceedings of the International Conference on Data Mining, Lake Buena Vista, FL, USA, 22–24 April 2004; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2004; pp. 522–526. [CrossRef]
47. Vaidya, J.; Clifton, C.; Kantarcioglu, M.; Patterson, A.S. Privacy-Preserving Decision Trees over Vertically Partitioned Data. *ACM Trans. Knowl. Discov. Data* **2008**, *2*, 1–27. [CrossRef]
48. Wright, R.; Yang, Z. Privacy-Preserving Bayesian Network Structure Computation on Distributed Heterogeneous Data. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; Association for Computing Machinery: New York, NY, USA, 2004; pp. 713–718. [CrossRef]
49. Vaidya, J.; Clifton, C. Privacy-Preserving k-Means Clustering over Vertically Partitioned Data. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–27 August 2003; Association for Computing Machinery: New York, NY, USA, 2003; pp. 206–215. [CrossRef]
50. Jagannathan, G.; Wright, R.N. Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL, USA, 21–24 August 2005; Association for Computing Machinery: New York, NY, USA, 2005; pp. 593–599. [CrossRef]
51. Sheikhalishahi, M.; Martinelli, F. Privacy preserving clustering over horizontal and vertical partitioned data. In *IEEE Symposium on Computers and Communications*; IEEE Computer Society: Washington, DC, USA, 2017; pp. 1237–1244. [CrossRef]
52. Fung, B.C.M.; Trojer, T.; Hung, P.C.K.; Xiong, L.; Al-Hussaeni, K.; Dssouli, R. Service-Oriented Architecture for High-Dimensional Private Data Mashup. *IEEE Trans. Serv. Comput.* **2012**, *5*, 373–386. [CrossRef]
53. Meyerson, A.; Williams, R. On the Complexity of Optimal K-Anonymity. In Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '04), Paris, France, 14–16 June 2004; Association for Computing Machinery: New York, NY, USA, 2004; pp. 223–228. [CrossRef]
54. Fung, B.; Wang, K.; Yu, P. Top-down specialization for information and privacy preservation. In Proceedings of the 21st International Conference on Data Engineering, Washington, DC, USA, 5–8 April 2005; pp. 205–216. [CrossRef]
55. Cárdenas-Robledo, L.A.; Peña-Ayala, A. Ubiquitous learning: A systematic review. *Telemat. Inform.* **2018**, *35*, 1097–1132. [CrossRef]
56. Chango, W.; Cerezo, R.; Romero, C. Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Comput. Electr. Eng.* **2021**, *89*, 106908. [CrossRef]

57. Waheed, H.; Hassan, S.U.; Aljohani, N.R.; Hardman, J.; Alelyani, S.; Nawaz, R. Predicting academic performance of students from VLE big data using deep learning models. *Comput. Hum. Behav.* **2020**, *104*, 106189. [CrossRef]
58. Zafra, A.; Romero, C.; Ventura, S. Multiple instance learning for classifying students in learning management systems. *Expert Syst. Appl.* **2011**, *38*, 15020–15031. [CrossRef]
59. Sheth, A. Internet of Things to Smart IoT Through Semantic, Cognitive, and Perceptual Computing. *IEEE Intell. Syst.* **2016**, *31*, 108–112. [CrossRef]
60. Pardo, A.; Siemens, G. Ethical and privacy principles for learning analytics. *Br. J. Educ. Technol.* **2014**, *45*, 438–450. [CrossRef]
61. Hoel, T.; Chen, W. Privacy-driven Design of Learning Analytics Applications—Exploring the Design Space of Solutions for Data Sharing and Interoperability. *J. Learn. Anal.* **2016**, *3*, 139–158. [CrossRef]
62. Kuzilek, J.; Hlosta, M.; Zdrahal, Z. Open University Learning Analytics dataset. *Sci. Data* **2017**, *4*, 170171. [CrossRef]
63. Vidal, V.M.P.; Casanova, M.A.; Cardoso, D.S. Incremental Maintenance of RDF Views of Relational Data. In Proceedings of the On the Move to Meaningful Internet Systems Conference, Rhodes, Greece, 21–25 October 2019; Meersman, R., Panetto, H., Dillon, T., Eder, J., Bellahsene, Z., Ritter, N., De Leenheer, P., Dou, D., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8185, pp. 572–587. [CrossRef]
64. Gharehchopogh, F.S.; Arjang, H. A Survey and Taxonomy of Leader Election Algorithms in Distributed Systems. *Indian J. Sci. Technol.* **2014**, *7*, 815–830. [CrossRef]
65. Mansour, E.; Sambra, A.V.; Hawke, S.; Zereba, M.; Capadisli, S.; Ghanem, A.; Aboulnaga, A.; Berners-Lee, T. A Demonstration of the Solid Platform for Social Web Applications. In Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, QC, Canada, 11–15 April 2016; pp. 223–226. [CrossRef]
66. Liu, G.Z.; Hwang, G.J. A key step to understanding paradigm shifts in e-learning: Towards context-aware ubiquitous learning. *Br. J. Educ. Technol.* **2010**, *41*, E1–E9. [CrossRef]
67. Escamilla-Ambrosio, P.; Rodríguez-Mota, A.; Aguirre-Anaya, E.; Acosta-Bermejo, R.; Salinas-Rosales, M. Distributing Computing in the Internet of Things: Cloud, Fog and Edge Computing Overview. In *Studies in Computational Intelligence*; Maldonado, Y., Trujillo, L., Schütze, O., Riccardi, A., Vasile, M., Eds.; Springer: Cham, Switzerland, 2018; Volume 731. [CrossRef]
68. Li, H.; Guo, F.; Zhang, W.; Wang, J.; Xing, J. (a,k)-Anonymous Scheme for Privacy-Preserving Data Collection in IoT-based Healthcare Services Systems. *J. Med. Syst.* **2018**, *42*, 56. [CrossRef] [PubMed]
69. Jara, A.J.; Olivieri, A.C.; Bocchi, Y.; Jung, M.; Kastner, W.; Skarmeta, A.F. Semantic Web of Things: An Analysis of the Application Semantics for the IoT Moving towards the IoT Convergence. *Int. J. Web Grid Serv.* **2014**, *10*, 244–272. [CrossRef]
70. Zamfiroiu, A.; Iancu, B.; Boja, C.; Georgescu, T.M.; Cartas, C.; Popa, M.; Toma, C.V. IoT Communication Security Issues for Companies: Challenges, Protocols and The Web of Data. *Proc. Int. Conf. Bus. Excell.* **2020**, *14*, 1109–1120. [CrossRef]
71. Hameed, S.S.; Hassan, W.H.; Latiff, L.A.; Ghabban, F. A systematic review of security and privacy issues in the internet of medical things; The role of machine learning approaches. *Peerj Comput. Sci.* **2021**, *7*, e414. [CrossRef]
72. Parikh, S.; Dave, D.; Patel, R.; Doshi, N. Security and Privacy Issues in Cloud, Fog and Edge Computing. *Procedia Comput. Sci.* **2019**, *160*, 734–739. [CrossRef]